Documentation for the iPool Extract API

1 Introduction

Aim of the iPool Extract API is semantic enrichment of texts. The enrichment is based on a REST API that can be used with a valid API key.

2 Linguistic enrichment

This documentation describes the functionality of entity and keyword enrichment within texts. Entities and keywords are initially recognized and returned as JSON. Return values are based on token, lemma and weighting.

2.1 Entities

The following types of proper nouns are recognized.

	description	examples
personal names	full personal names with first, middle and last names	Angela Merkel, Nikolaus Pöhlchen, Karl-Theodor zu Guttenberg
companies, organisations	known and unknown companies and organizations	Porsche AG, Bachmann GmbH, Lufthansa, SWU
geographical entities	places, cities, countries, rivers, mountains	Berlin, Deutschland, Alpen, Erbach, Altheim
events	large german and international events will be detected	Bambi Verleihung, MTV Music Award, Cebit
products	products will be detected	iPhone, iPad, Galaxy Tab

The named entity recognition works with grammatical rules, especially previously unknown proper names that were not listed on internal lists are detected automatically. Proper names are mapped via acronyms, the acronym letters are preferred, such as „Deutsche Lebens-Rettungs-Gesellschaft e. V.“ = DLRG.

2.2 Keywords

The automatic tag assignment takes place on basis of the text, especially nouns will be used. Entities won't be tagged again, because entity recognition takes place before keyword extraction. The keywords will be brought to a normal form and added to the content. The selection is performed based on article relevance and general keyword relevance. This leads, among other things, that buzzwords like "house" or "city" does not come into consideration, however, "bus accident" or "abuse of office" will have a good chance.

3 Linguistics-API (Enricher)

In the following sections the Extract API endpoint for linguistic analysis will be explained. The returned format is JSON.

3.1 Function: Annotation of a document

function name	/extract	
REST request	POST	
parameters	title	string
	subtitle	string
	content	string
returns	200	OK → JSON result
	403	forbidden
	400	invalid request data

At least one of the parameters title, subtitle or content has to be set to get results. For better recognition and relevance ranking of entities we recommend to pass the fields title and subtitle to the enrichment process.

3.1.1 Return Values

The returned JSON mainly includes three information categories: Entities and keywords recognized by linguistic component and an associated relevance value. The relevance value can be used e.g. to control which entities could be used within tag clouds or for relevant topic pages.

weight: Refers to the relevance of the entity and keyword.

Entities in title and subtitle will be weighted higher than e.g. the actual occurrence of the entity in the body of the document. In addition, the position of the entity within the article will be considered. An entity at the end of the article is less important than the entity at the beginning of the article.

token: The token returns the keyword, as it was found within the analyzed text.

lemma: Is the base form of the keyword.

3.1.2 Coordinates for geo entities

The return value includes also the geo coordinates for the most relevant geo within title and subtitle. The content parameter is not considered for extracting coordinates. The attribute location within the response delivers corresponding latitude and longitude.

location: Geo coordinates for the most relevant geo in title/subtitle

3.1.3 Example

Semantic enrichment of

http://www.welt.de/politik/deutschland/article140263646/Ruestungsfirmen-aus-China-sind-besonders-korrupt.html

will result in

location	{ "location": "35.0,105.0",
events	"events": [],
geos	"geos": [{ "lemma": "China", "token": ["China"], "weight": 13.482748 }, { "lemma": "Deutschland", "token": ["Deutschland"], "weight": 5.706281 }],

keywords	"keywords": [
	{
	"lemma": "Rüstungsunternehmen",
	"token": [
	"Rüstungsunternehmen"
	],
	"weight": 29.539875
	},
	{
	"lemma": "Korruption",
	"token": [
	"Korruption"
	],
	"weight": 162.0584
	},
	{
	"lemma": "Whistleblower",
	"token": [
	"Whistleblower"
	],
	"weight": 10.917133
	}
	],

orgs


```
"orgs": [
  {
    "lemma": "Diehl Stiftung",
    "token": [
      "Diehl Stiftung & Co. KG"
    ],
    "weight": 10
  },
  {
    "lemma": "MTU Aero Engines",
    "token": [
      "MTU Aero Engines Holding AG"
    ],
    "weight": 10
  },
  {
    "lemma": "Thyssen Krupp",
    "token": [
      "Thyssen Krupp AG"
    ],
    "weight": 10
  },
  {
    "lemma": "Bundesregierung",
    "token": [
      "Bundesregierung"
    ],
    "weight": 29.044529
  },
  {
    "lemma": "Krauss-Maffei Wegmann",
    "token": [
      "Krauss - Maffei Wegmann"
    ],
    "weight": 30
  },
  {
    "lemma": "Rheinmetall",
    "token": [
      "Rheinmetall"
    ],
    "weight": 8.749631
  },
  {
    "lemma": "ThyssenKrupp",
    "token": [
      "ThyssenKrupp"
    ],
    "weight": 8.470953
  }
],
```

persons	"persons": [
	{
	"lemma": "Edda Müller",
	"token": [
	"Edda Müller"
	],
	"weight": 3.0389266
	}
	],

products	"products": [] }